CARL VON OSSIETZKY universität OLDENBURG

On the Accuracy of Online Geocoders

Dirk Ahlers
OFFIS, Oldenburg
Susanne Boll
University of Oldenburg

OFFIS
INSTITUTE FOR INFORMATION TECHNOLOGY

01.04.2009
Geoinformatik 2009
Osnabrück
52° 16' 22" N 8° 2' 30"E

Geographic Web Information Retrieval

- Matching of Web information to the real world
- Semantic retrieval of location information
- Identification, extraction and processing of geographic references

Motivation

- Mobile users require local information
- Up to 20% location-related queries
- Location references in up to 20% of Web pages
- Richer information than most commercial databases
- WWW as source of location-based information

Location information on the Web

- Geospatial information on the Web is …
  - Hidden in Web pages
  - Scattered
  - Not explicitly structured

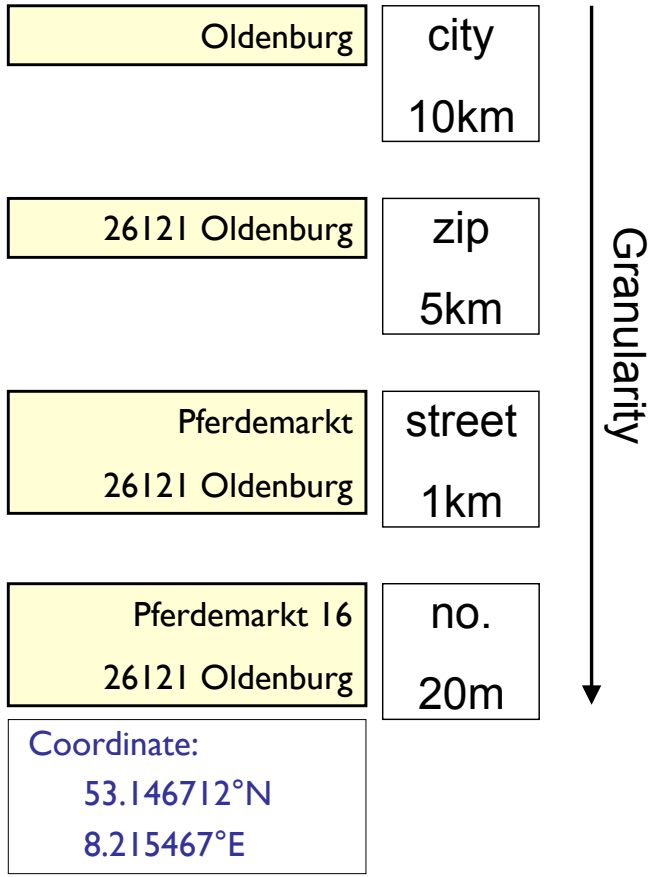
- Metadata
  - Geographic coordinates
- Content
  - Addresses

```
<META name="geo.position"
      content="53.1467;8.2154">
```

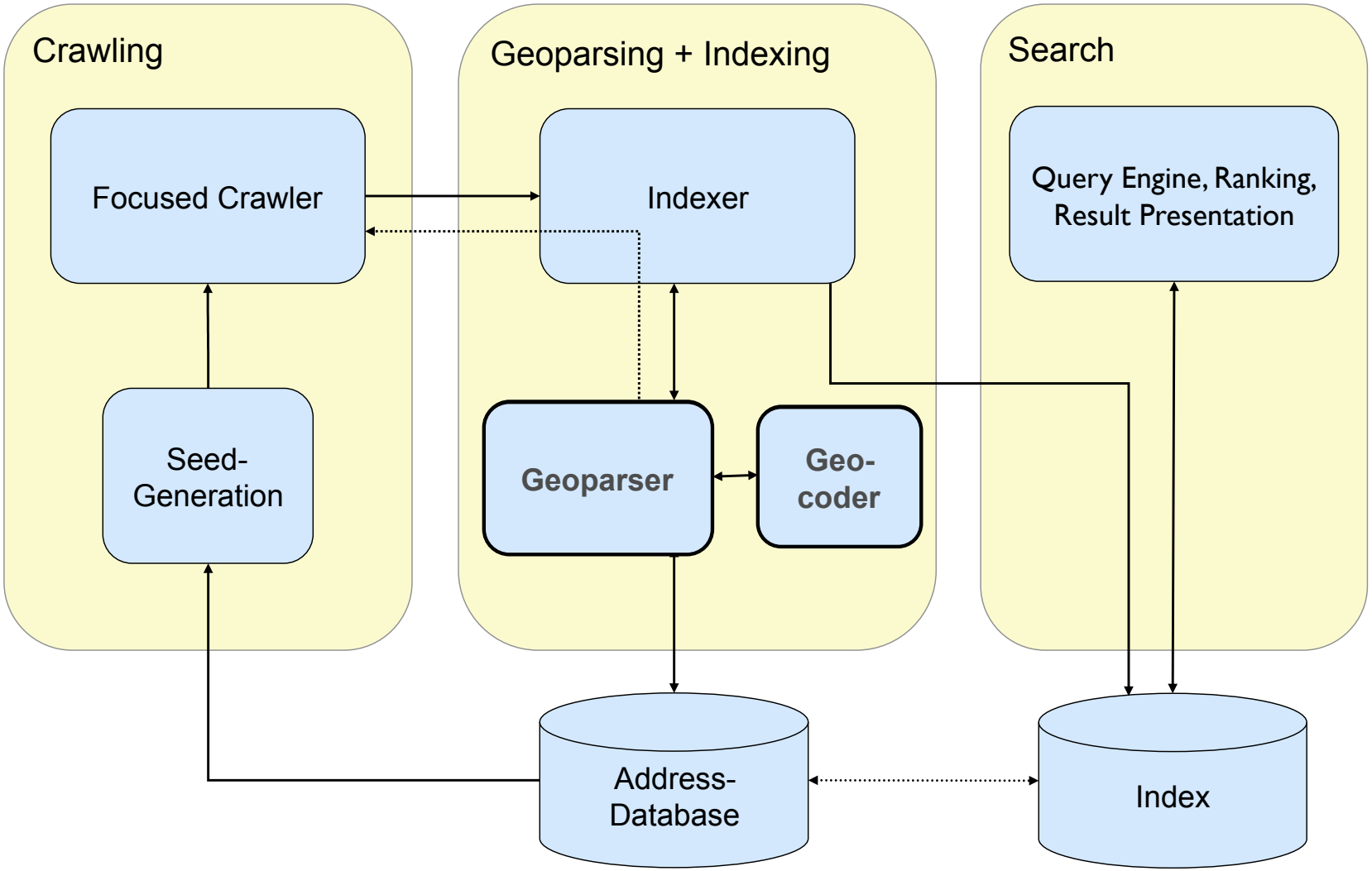
Pferdemarkt 15
26121 Oldenburg

Granularity of geographic references

- Additional location terms increase granularity
- Highest increase with addition of house number
- Full address allows ‘exact’ geocoding to an individual building



Geospatial Search Engine Architecture

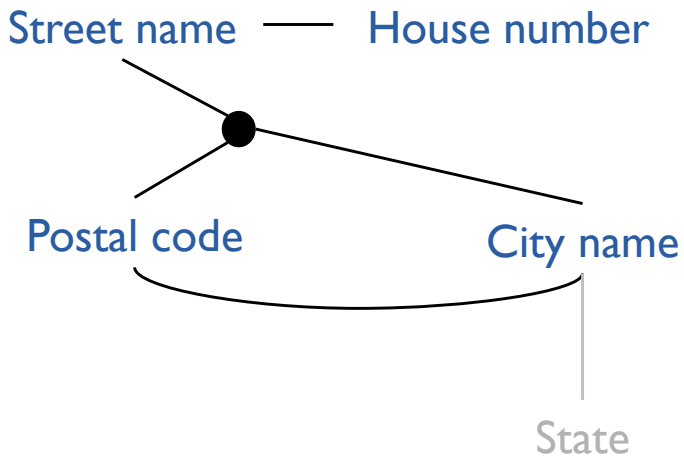


Address geoparsing

- Identification and annotation of location references in unstructured Web pages
- Database-backed extraction
- Address database as gazetteer
- Heuristic rule-based extraction
- Disambiguation of terms
- Validating geoparser

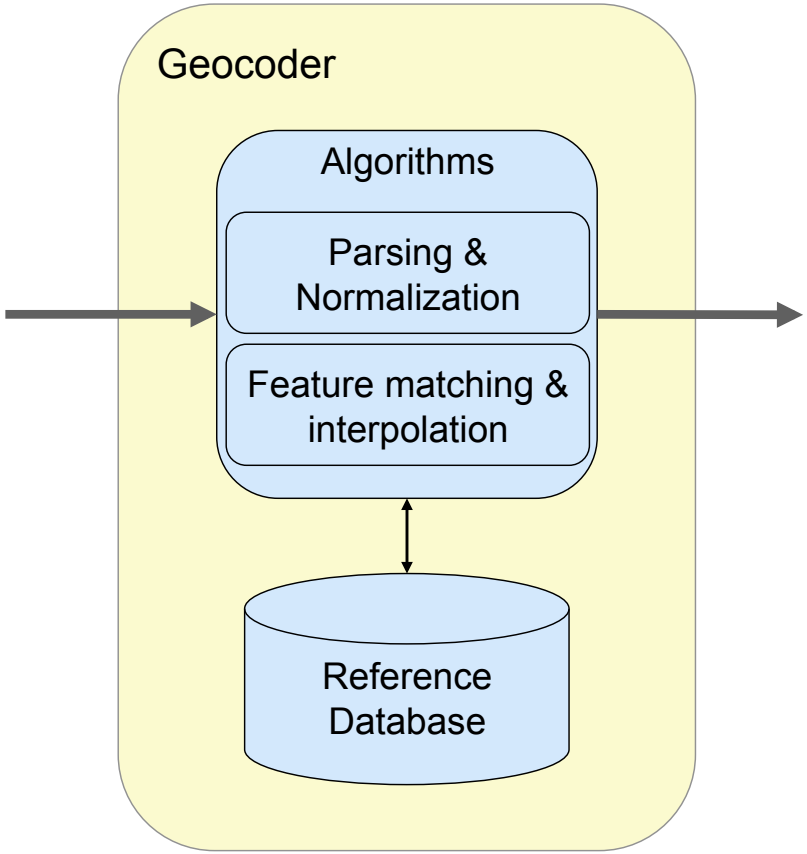
Pferdemarkt 15
26121 Oldenburg

```
<ul>
  <li class="b"><span
class="rot">Landesbibliothek</span><br><span
class="blau2">Oldenburg</span></li>
  <li>Pferdemarkt 15</li>
  <li>26121 Oldenburg</li>
  <li>Tel. 0441-799-2800</li>
  <li>Fax. 0441-799-2865</li>
  <li>...
</ul>
```



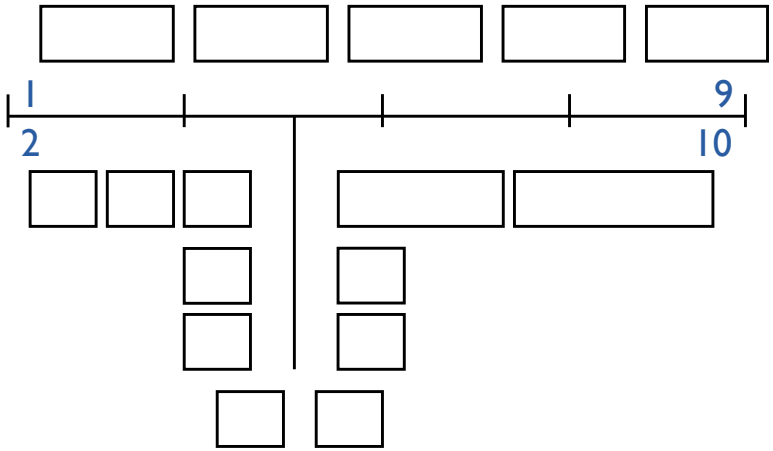
Geocoding

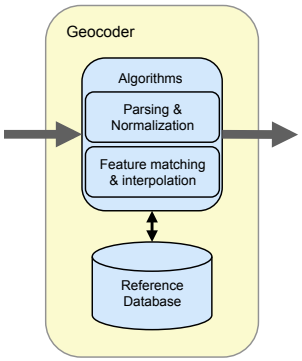
- Conversion of textual place descriptions to geographic coordinates
- Geocoder
  - Functions and heuristics
  - Reference database



Error Types

- Geocoding
  - Line simplification
  - Interpolation errors
  - Inconsistent normalization
  - Scrambled numbering schemes
  - Offset/Inset variations
- Reference database
  - Incomplete coverage
  - Outdated entries
  - Missing streets / street parts
  - Nonmodeled gaps





Challenge

- Only two freely available geocoders
  - Geocoder Web services from Yahoo! and Google
  - Good coverage at address-level
- Results differ
  - No majority vote
- Little metadata annotated
- No common ground truth available

Geocoders as blackboxes?

- Database and algorithms unknown
- But: Metadata on results
  - Accuracy measure, but no confidence

Google	Yahoo!	Description of accuracy level
0	warning or error	Unknown location.
1	country	Country
2	state	Region (state, province, prefecture, etc.)
3		Sub-region (county, municipality, etc.)
4	city	Town (city, village)
5	zip	Post code
	zip+2	Post code + 2 digits (US)
	zip+4	Post code + 4 digits (US)
6	street	Street
7		Intersection
8	address	Address
9		Premise (building name, property name, etc.)

Correction by metadata

- Simple approach
  - Query both services
  - Use metadata to select result with better accuracy

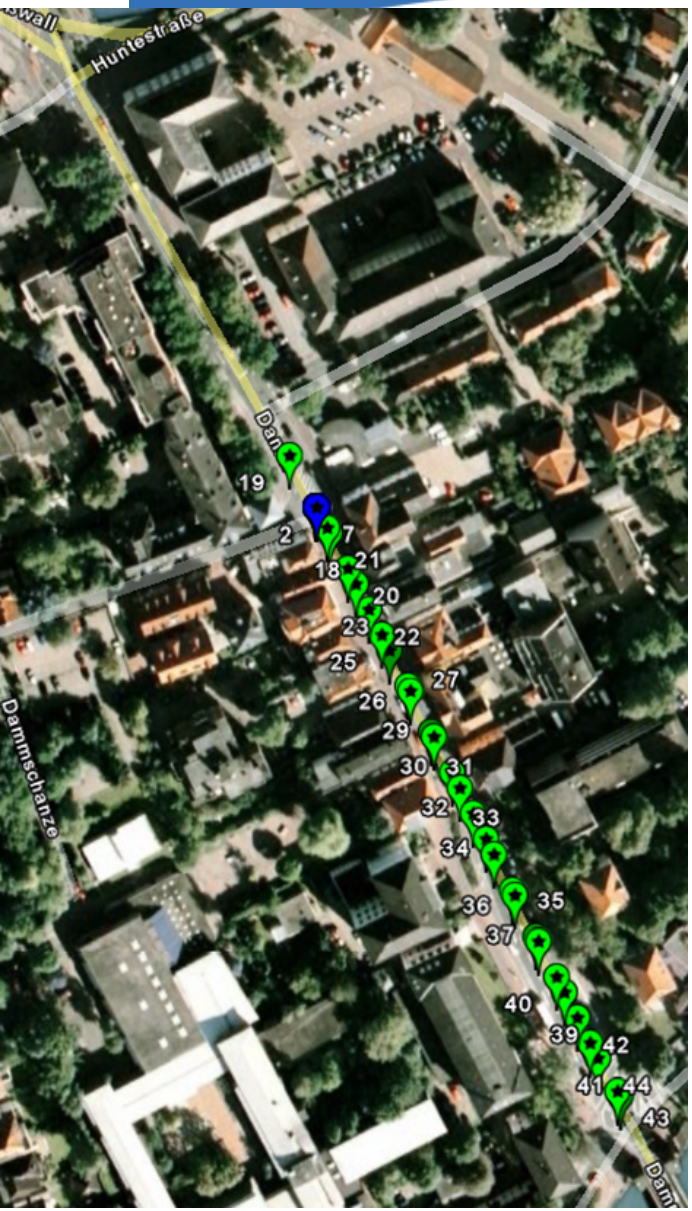
Geocoder1	Geocoder2	Result
street	street	Granularity is too low. Use the preferred geocoder, an average, or mark for manual geocoding
address	street	Use geocoder 2
street	address	Use geocoder 1
address	address	combine results?

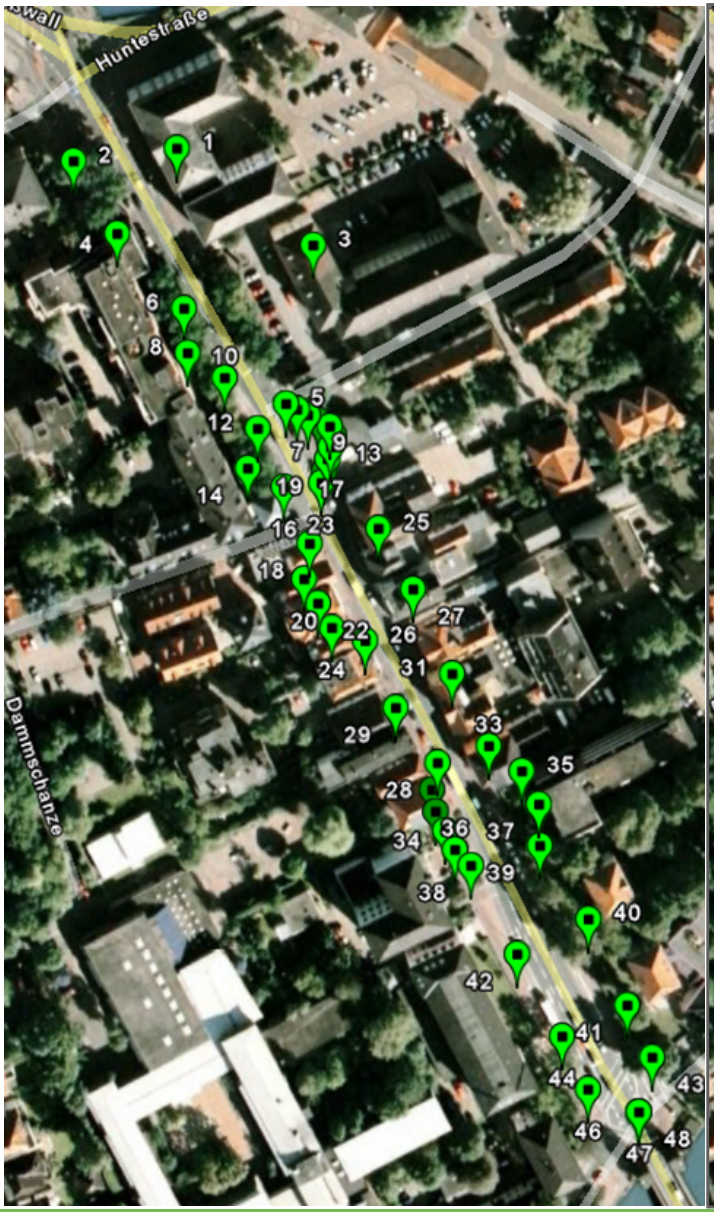
Correction by geographical analysis

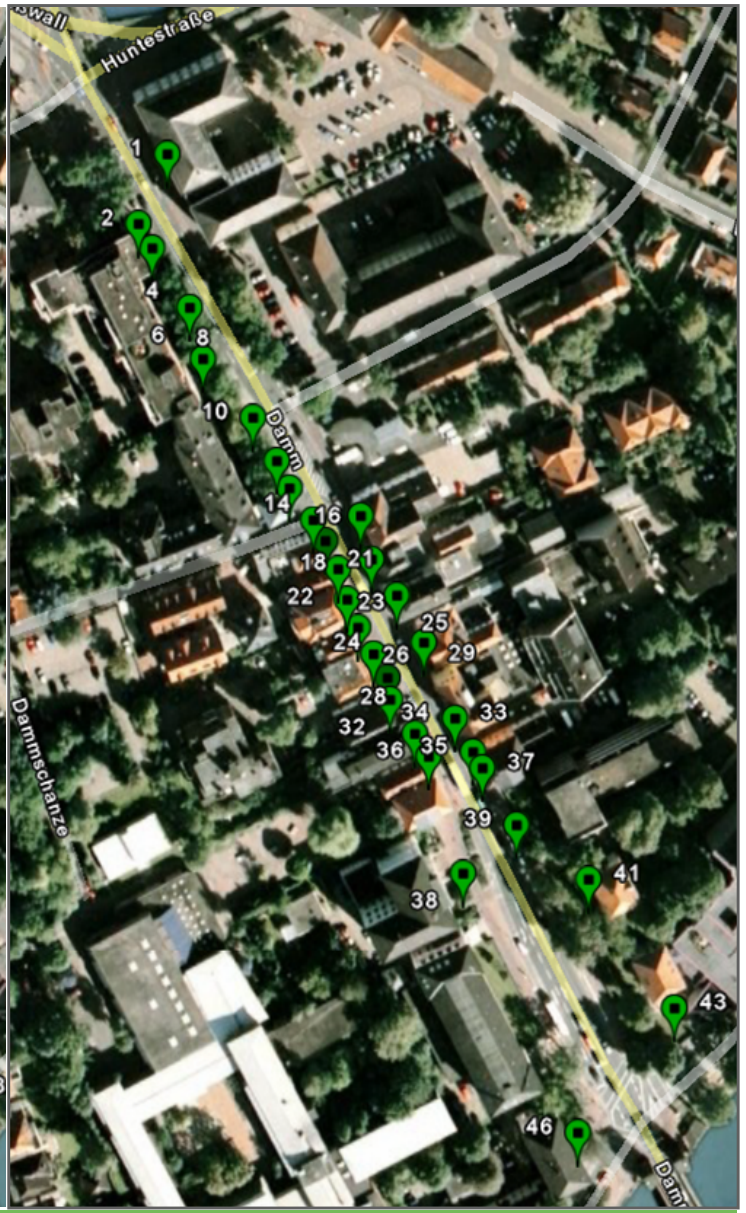
- Mismatched results with address-level accuracy indicator need to be assessed
- Within small threshold causeable by offsets
  - Average is often sufficient
- Larger differences in position
  - Surrounding house numbers are checked
  - Use of position, outliers, repetition, clusters, smoothness

Example

Example


Huntestraße
Dammschanze


Huntestraße
Dammschanze


Huntestraße
Dammschanze

Results

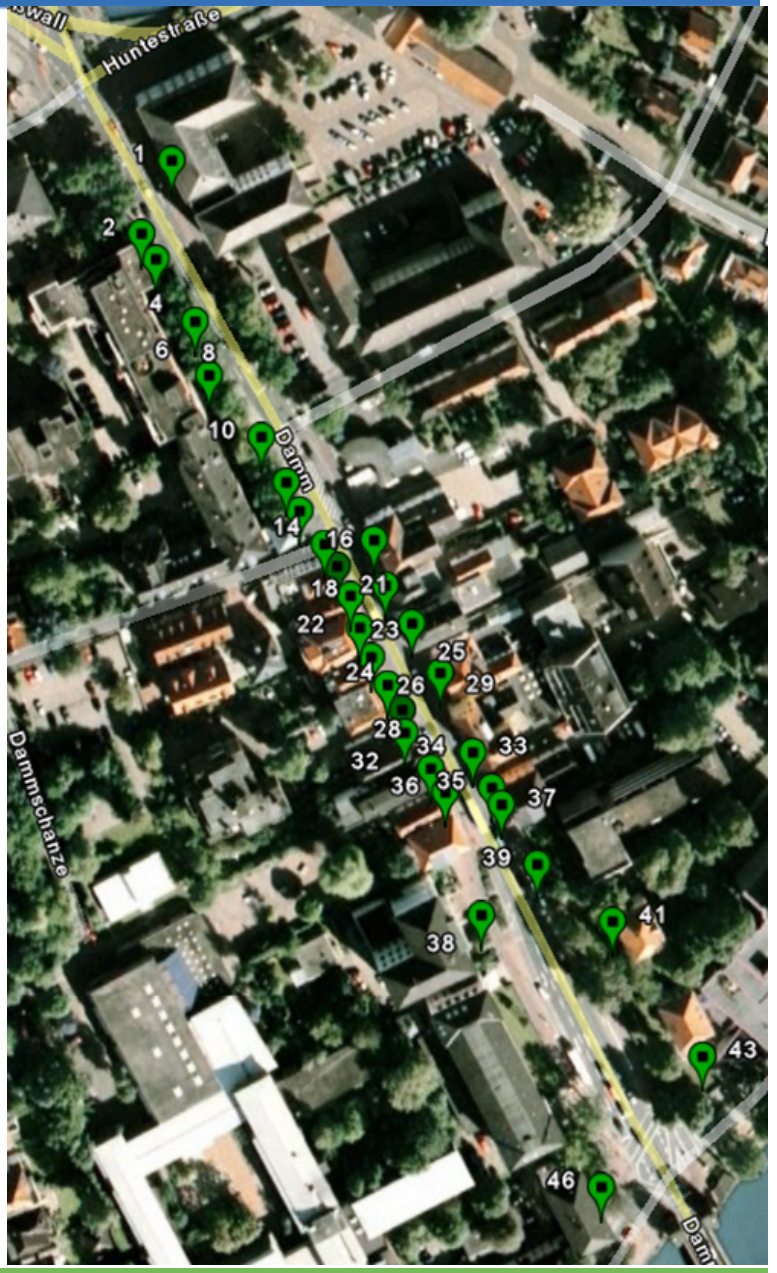
geocoder	*mean*	σ	*min*	*max*
google	179,2	512,3	6,2	2349,1
yahoo	225,8	549,0	0,0	2288,6
corr	169,4	490,4	0,0	2285,6
Artillerieweg_google	25,6	11,8	7,1	53,9
Artillerieweg_yahoo	28,0	20,2	6,0	77,9
Artillerieweg_corr	18,9	14,1	0,0	57,7
Damm_google	50,2	33,5	8,1	137,6
Damm_yahoo	21,5	13,7	0,0	49,8
Damm_corr	24,7	15,8	0,0	51,4
Staustrasse_google	153,6	362,6	6,2	1413,9
Staustrasse_yahoo	179,9	231,0	0,0	696,4
Staustrasse_corr	73,7	111,4	0,0	351,2

Deviation from manually determined ground truth in m

Conclusion

- High granularity is freely available
- Conflation of two sources possible
- Geocoder accuracy can be improved
- Some cross-source errors cannot currently be detected

- Future work
  - Improvements on remaining error cases
  - Use of third source
  - OSM house numbers improving



Q&A

Dirk Ahlers
OFFIS Oldenburg
53° 8′ 55.9″ N 8° 12′ 0.43″ E
ahlers@offis.de